

## Accepted Manuscript

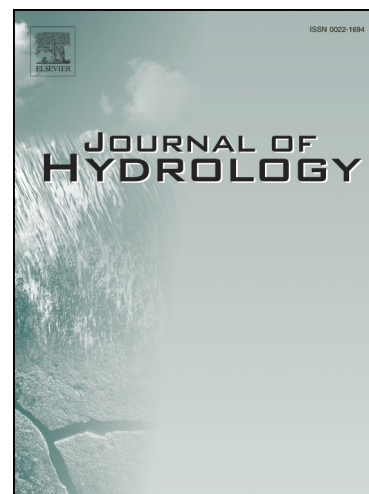
Spatio-Temporal Modelling of Rainfall in the Murray-Darling Basin

Gen Nowak, A.H. Welsh, T.J. O'Neill, Lingbing Feng

PII: S0022-1694(17)30784-9  
DOI: <https://doi.org/10.1016/j.jhydrol.2017.11.021>  
Reference: HYDROL 22379

To appear in: *Journal of Hydrology*

Received Date: 13 March 2017  
Revised Date: 26 September 2017  
Accepted Date: 11 November 2017



Please cite this article as: Nowak, G., Welsh, A.H., O'Neill, T.J., Feng, L., Spatio-Temporal Modelling of Rainfall in the Murray-Darling Basin, *Journal of Hydrology* (2017), doi: <https://doi.org/10.1016/j.jhydrol.2017.11.021>

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

# Spatio-Temporal Modelling of Rainfall in the Murray-Darling Basin

Gen Nowak<sup>a,\*</sup>, A. H. Welsh<sup>b</sup>, T. J. O'Neill<sup>c</sup>, Lingbing Feng<sup>d</sup>

<sup>a</sup>*Research School of Finance, Actuarial Studies and Statistics, ANU College of Business and Economics,  
The Australian National University, Acton, ACT, 2601, Australia*

<sup>b</sup>*Mathematical Sciences Institute, ANU College of Physical and Mathematical Sciences, The Australian  
National University, Acton, ACT, 2601, Australia*

<sup>c</sup>*Bond Business School, Bond University, Robina, QLD, 4226, Australia*

<sup>d</sup>*International Institute for Financial Studies, Jiangxi University of Finance and Economics, Nanchang,  
Jiangxi, 33013, China*

---

## Abstract

The Murray-Darling Basin (MDB) is a large geographical region in southeastern Australia that contains many rivers and creeks, including Australia's three longest rivers, the Murray, the Murrumbidgee and the Darling. Understanding rainfall patterns in the MDB is very important due the significant impact major events such as droughts and floods have on agricultural and resource productivity. We propose a model for modelling a set of monthly rainfall data obtained from stations in the MDB and for producing predictions in both the spatial and temporal dimensions. The model is a hierarchical spatio-temporal model fitted to geographical data that utilises both deterministic and data-derived components. Specifically, rainfall data at a given location are modelled as a linear combination of these deterministic and data-derived components. A key advantage of the model is that it is fitted in a step-by-step fashion, enabling appropriate empirical choices to be made at each step.

**Keywords:** Spatio-Temporal, Rainfall, Murray-Darling Basin, Prediction, Bootstrap

---



---

\*Corresponding author

Email address: [gen.nowak@anu.edu.au](mailto:gen.nowak@anu.edu.au) (Gen Nowak)

## 1. Introduction

Climate change is a globally recognised issue with far reaching ecological, environmental and agricultural consequences. For example, in Australia, rising sea temperatures have damaged the Great Barrier Reef and the Millenium Drought severely affected agriculture in much of southern Australia. One particular region of Australia that has been adversely impacted by the increasing frequency of extreme weather events is the Murray-Darling Basin (MDB). The MDB, displayed in Figure 1, is a large geographical region of over 1 million square kilometres in southeastern Australia. The region spans four states and contains many rivers and wetlands. It is the most significant agricultural area in Australia, containing more than half the nation's irrigated farms that are responsible for much of the nation's irrigated produce. As such, water management in the MDB is a very important issue.

The main focus of this paper is to model and predict rainfall in the MDB. A long-term goal is to model water levels in the MDB through its relationship with rainfall. Developing a better understanding of how rainfall fluctuates and, consequently, how this affects river levels is critical for water management in the MDB. Some recent, relevant work on the analysis of rainfall data in the MDB can be found in Potter et al. (2010), Smith and Chandler (2010) and Feng et al. (2014). Potter et al. (2010) analysed times series of annual rainfall and runoff to detect trends and step changes in the data, Smith and Chandler (2010) assessed a number of rainfall projection models in order to identify a subset of better performing models and Feng et al. (2014) developed a method for imputing spatio-temporal rainfall data based on spatial correlation and cross-validation.

The data we are modelling are spatio-temporal in nature, consisting of rainfall measurements taken over time at various locations within the MDB. Spatio-temporal data arise in many situations, ranging from climatology, epidemiology, geology to environmental health. An overview of spatio-temporal data and modelling of spatio-temporal data can be found in Cressie and Wikle (2011) and Banerjee et al. (2015). Some recent work in spatio-temporal

28 modelling includes Gryparis et al. (2007), Bogaert et al. (2009), Eckert et al. (2010), Holly  
 29 et al. (2010), Fonseca and Steel (2011) and Lowe et al. (2011), with specific applications to  
 30 rainfall data given in Allcroft and Glasbey (2003), Carrera-Hernández and Gaskin (2007)  
 31 and Sigrist et al. (2012). Further studies in spatio-temporal analysis that focus on aspects  
 32 such as anisotropy and extremes are given in Rodrigues et al. (2015), Zhao (2015), Comas  
 33 et al. (2015), Lovino et al. (2014) and Ghosh and Mallick (2011). Much of the current  
 34 literature in spatio-temporal modelling adopts a Bayesian approach, as the hierarchical  
 35 nature of these models naturally lends itself to this framework.

36 While Bayesian models are popular and useful for spatio-temporal data, they are of-  
 37 ten computationally intensive and can sometimes be difficult to interpret, especially when  
 38 dealing with predictions. We propose a non-Bayesian hierarchical model for the spatio-  
 39 temporal rainfall data in the MDB. Our approach models the time series at each spatial  
 40 location as a linear combination of basis functions. These basis functions represent tem-  
 41 poral patterns or features that are shared among the spatial locations. We can account for  
 42 variability among spatial locations by allowing the coefficients of the basis functions to be  
 43 spatially dependent.

44 Our methodology involves four key novel aspects. First, we fit a hierarchical model in  
 45 a step-by-step procedure using a frequentist approach, rather than a Bayesian approach.  
 46 This enables diagnostics to be performed, and empirical choices to be made, at each step.  
 47 Second, we established a new method for deriving the basis functions that incorporates both  
 48 deterministic and data-derived components. Third, we developed a block bootstrap method  
 49 for producing parameter estimate standard errors that maintains structural relationships  
 50 present in the data. Last, the model produces predictions, both in the future and at  
 51 unobserved spatial locations, in a natural, intuitive way.

52 The paper is organised as follows: Section 2 describes the monthly rainfall data that  
 53 we are analysing, Section 3 details the proposed model, Section 4 outlines the model fitting  
 54 and parameter estimation procedure and Section 5 presents the predictive performance of

the model.

## 2. High-Quality Monthly Rainfall Data

The monthly rainfall data were obtained from a network of weather stations at which high-quality data are available (<http://www.bom.gov.au/climate/change/hqsites/>). As our study focused on the MDB, we selected among this network of stations those that fell within the MDB. There were a total of 78 such stations, which are displayed in Figure 1. Following the Bureau of Meteorology, we call the data set the high-quality monthly rainfall (HQMR) data.

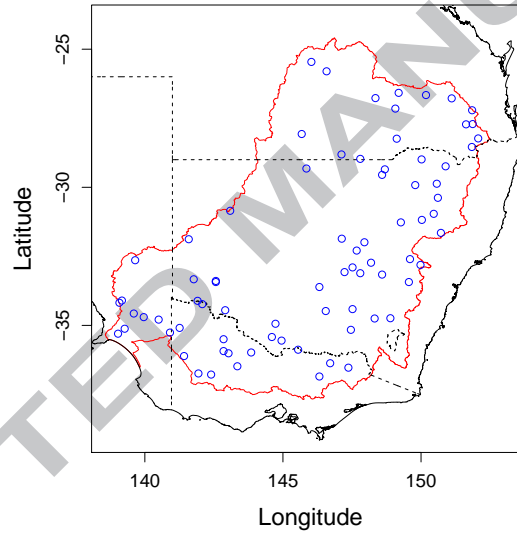


Figure 1: The weather stations within the MDB (outlined in red) for which the HQMR data were available.

For each station, we have the station's latitude, longitude, elevation and which of the three climatic regimes defined by Connell and Grafton (2011, Figure 1.4) it belongs to. We also have monthly rainfall measurements that have been recorded at each station over many years. Displayed in Figure 2 are the span of months for which each station recorded monthly rainfall readings. We see that the range of dates varies among stations, with

68 readings ranging from as early as January 1868 to as recently as February 2011. Further,  
 69 prior to the early 1900s, due to the varying starting dates, there are much missing data.  
 70 In order to maximise the completeness and accuracy of our data, we focused our analysis  
 71 on a subset of the data, ranging from the latest date at which a station began recording  
 72 measurements (January 1923) to the earliest date at which a station stopped recording  
 73 measurements (February 2005). This period is indicated by the red bars in Figure 2.

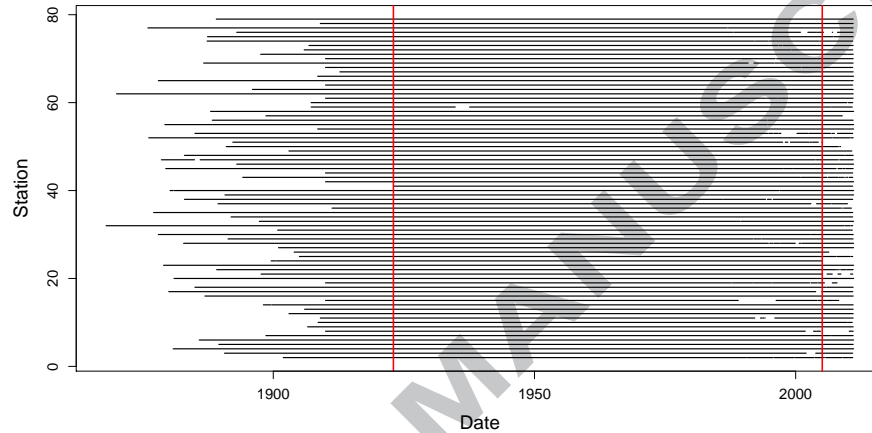


Figure 2: The months for which the HQMR measurements were recorded for stations in the MDB. The red bars indicate the time period from January 1923 and February 2005.

74 Some exploratory plots are displayed in Figures 3 to 5. A histogram of the HQMR  
 75 measurements over all stations and months is given in Figure 3. Due to the frequent pe-  
 76 riods of drought in Australia, there is a very high frequency of months where very little  
 77 rainfall was recorded. In addition, the measurements are highly positively skewed, indicat-  
 78 ing that it may be necessary to transform the data. Although a log transformation is often  
 79 used for data of this nature, the presence of zero measurements is problematic. Therefore,  
 80 we applied a cube-root transformation to the data and the corresponding histogram is  
 81 also displayed in Figure 3. The cube-root transformation, which is also commonly used  
 82 on rainfall data (Feng et al., 2014), effectively reduces the skewness. In Figure 4, both  
 83 the untransformed and transformed HQMR measurements for each month, averaged over

stations, are plotted against time. While these plots do not seem to show any obvious long-term temporal trends, there are likely to be some seasonal patterns. The cube-root transformation again has reduced the spread and skewness in these station-averaged HQMR measurements. In Figure 5, the transformed HQMR measurements for each station, averaged over time, are plotted spatially as a bubble plot. We also fitted a smooth surface to these time-averaged HQMR measurements using barycentric interpolation (linear interpolation within the triangles bounded by the data points). A 3D plot and a contour plot of the fitted surface are included in Figure 5. In terms of any observed spatial trends, the rainfall tends to increase along the southern and eastern boundaries of the MDB, i.e., as we approach the Great Dividing Range. This likely indicates that the spatial process is anisotropic and we may need to take this into account when modelling the HQMR data. This may be explained by a spatial-temporal trend in the sense that the residuals, after removing the trend, show only isotropic spatial correlation. Note that all further analyses performed in this paper were applied to the cube-root transformed data.

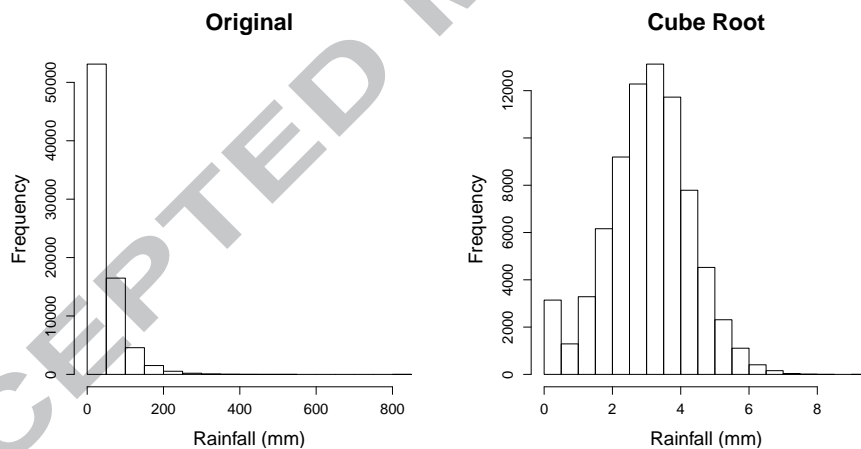


Figure 3: Histogram of all the HQMR measurements for all stations and months, for the untransformed data (left) and the cube-root transformed data (right).

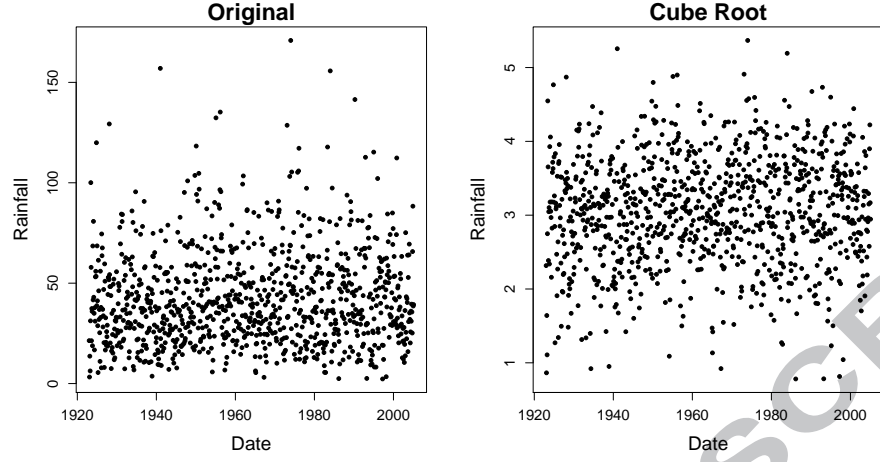


Figure 4: Scatterplot of the HQMR measurements, averaged over stations, against time for the untransformed data (left) and the cube-root transformed data (right).

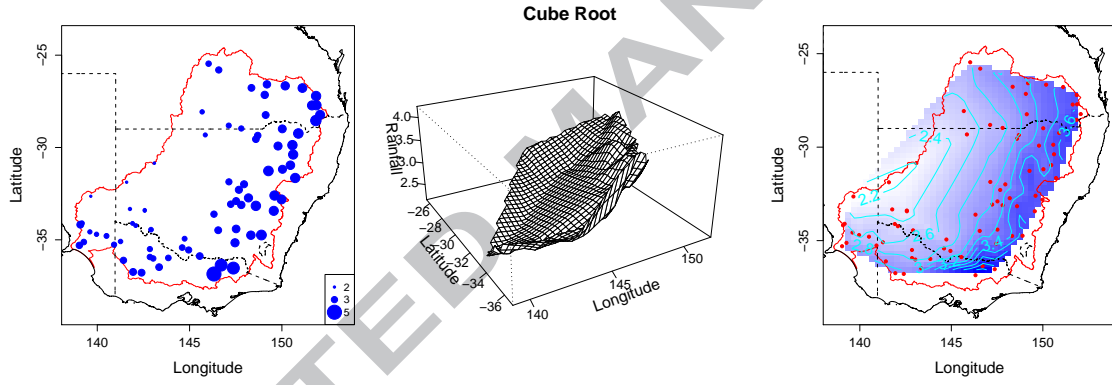


Figure 5: Cube-root transformed HQMR measurements, averaged over time, displayed spatially as a bubble plot (left). A fitted smooth surface is displayed as 3D plot (middle) and a contour plot (right).

### 3. Proposed Model

#### 3.1. Model

Initially, we will model the HQMR measurements using an approach similar to that employed by Lindstrom et al. (2011), Szpiro et al. (2010) and Fuentes et al. (2006). Let  $\{s_i\}_{i=1}^N$  denote the spatial locations of the  $N = 78$  stations. Let  $t = 1, \dots, T$  be a discrete



time index for month, spanning the period from January 1923 to February 2005 ( $T = 986$ ).

Letting  $Y(\mathbf{s}, t)$  denote the cube-root transformed HQMR measurement at any spatial location  $\mathbf{s}$  for month  $t$ , we use the following model for  $Y(\mathbf{s}, t)$ :

$$Y(\mathbf{s}, t) = \mu(\mathbf{s}, t) + e(\mathbf{s}, t),$$

where  $\mu(\mathbf{s}, t)$  denotes the mean spatio-temporal structure and  $e(\mathbf{s}, t)$  denotes the residual component. Therefore, for each station,  $i = 1, \dots, 78$ , and month,  $t = 1, \dots, 986$ , we have

$$Y(\mathbf{s}_i, t) = \mu(\mathbf{s}_i, t) + e(\mathbf{s}_i, t). \quad (1)$$

We will model  $Y(\mathbf{s}, t)$  hierarchically by proposing further models for both the mean spatio-temporal structure and the residual component.

Starting with the mean spatio-temporal structure, we model  $\mu(\mathbf{s}_i, t)$  as

$$\mu(\mathbf{s}_i, t) = \sum_{j=0}^J \beta_j(\mathbf{s}_i) f_j(t), \quad (2)$$

where  $f_0(t) \equiv 1$ ,  $\{f_j(t)\}_{j=1}^J$  are a set of smooth temporal basis functions and  $\{\beta_j(\mathbf{s}_i)\}_{j=0}^J$  are the corresponding spatially-varying coefficients. The number of basis functions  $J$  is generally small and details of their derivation are given in Section 4.1. The idea behind (2) is that the  $f_j(t)$  describe seasonal and long-term temporal trends that may be present in the data. For example,  $f_1(t)$  might describe the average temporal trend across all stations and subsequent basis functions may capture more subtle trends among individual stations. By incorporating spatially-varying coefficients for these temporal basis functions, we allow the temporal trends to differ among the stations. This is achieved by modelling  $\boldsymbol{\beta}_j = (\beta_j(\mathbf{s}_1), \dots, \beta_j(\mathbf{s}_N))^T$  as

$$\boldsymbol{\beta}_j \sim N(\mathbf{X}_j \boldsymbol{\alpha}_j, \boldsymbol{\Sigma}_{\beta_j}(\boldsymbol{\theta}_j)), \quad (3)$$

where  $\mathbf{X}_j$  is an  $N \times p_j$  matrix of spatial covariates,  $\boldsymbol{\alpha}_j$  is the corresponding  $p_j \times 1$  vector of coefficients and  $\boldsymbol{\Sigma}_{\beta_j}(\boldsymbol{\theta}_j)$  is an  $N \times N$  covariance matrix. The matrix  $\mathbf{X}_j$  depends on  $j$

and we can therefore fit a different set of spatial covariates for each  $j$ . Also, the  $\beta_j$  are assumed to be independent.

Moving onto the residual component, letting  $\mathbf{e}_t = (\mathbf{e}(\mathbf{s}_1, t), \dots, \mathbf{e}(\mathbf{s}_N, t))^T$ , we model  $\mathbf{e}_t$  as

$$\mathbf{e}_t \sim N(0, \Sigma_{\mathbf{e}_t}(\boldsymbol{\theta}_e)), \quad (4)$$

where  $\Sigma_{\mathbf{e}_t}(\boldsymbol{\theta}_e)$  is an  $N_t \times N_t$  covariance matrix, with  $N_t$  denoting the number of measurements at time  $t$ . If there are no missing data, then there will be a single  $N \times N$  covariance matrix,  $\Sigma_e(\boldsymbol{\theta}_e)$ , for all  $t$ . If there are missing data, the dimension of the covariance matrix can change with  $t$  and  $\Sigma_{\mathbf{e}_t}(\boldsymbol{\theta}_e)$  will be a sub-matrix of  $\Sigma_e(\boldsymbol{\theta}_e)$ . Any temporal structure present in the data is assumed to be captured by  $\mu(\mathbf{s}_i, t)$ , so the residuals  $\mathbf{e}_t$  are therefore assumed to be independent.

### 3.2. Parameters

From equations (1) to (4), the overall model that we are fitting is

$$Y(\mathbf{s}_i, t) = \sum_{j=0}^J \beta_j(\mathbf{s}_i) f_j(t) + e(\mathbf{s}_i, t), \quad (5)$$

where

$$\boldsymbol{\beta}_j \sim N(\mathbf{X}_j \boldsymbol{\alpha}_j, \Sigma_{\boldsymbol{\beta}_j}(\boldsymbol{\theta}_j)) \quad \text{and} \quad \mathbf{e}_t \sim N(0, \Sigma_{\mathbf{e}_t}(\boldsymbol{\theta}_e)). \quad (6)$$

Therefore, the parameters of the model that we need to estimate are:

- The intercept coefficients for  $f_0(t) = 1$ :  $\boldsymbol{\beta}_0 = (\beta_0(\mathbf{s}_1), \dots, \beta_0(\mathbf{s}_N))^T$ .
- The coefficients for each smooth temporal basis function  $f_j(t)$ :  $\boldsymbol{\beta}_j = (\beta_j(\mathbf{s}_1), \dots, \beta_j(\mathbf{s}_N))^T$  for  $j = 1, \dots, J$ .
- The coefficients for each set of spatial covariates  $\mathbf{X}_j$ :  $\boldsymbol{\alpha}_j = (\alpha_{j0}, \dots, \alpha_{jp_j})^T$  for  $j = 0, \dots, J$ .
- The parameters of the covariance function for each  $\boldsymbol{\beta}_j$ :  $\boldsymbol{\theta}_j$  for  $j = 0, \dots, J$ .

- The parameters of the covariance function for the residuals:  $\theta_e$ .

The parameters of covariance functions, i.e.,  $\theta_j$  for  $j = 0, \dots, J$  and  $\theta_e$ , will depend on the choice of the covariance function used. Determining the appropriate covariance functions will need to be based on the data and this is explored further in Section 4.2.

## 4. Model Fitting

### 4.1. Deriving the Temporal Basis Functions

An important step in the model specification process is to determine the smooth temporal basis functions that should be used for the data. The approach used by Szpiro et al. (2010) is to set the  $f_j(t)$ , for  $j = 1, \dots, J$ , to be smoothed versions of the first  $J$  left singular vectors from the singular value decomposition (SVD) of the  $T \times N$  data matrix  $\mathbf{Y}$ . Since  $\mathbf{Y}$  contains missing values, the SVD is calculated through an iterative procedure (described in Algorithm 1) that involves imputing the missing values. Note that the imputation of the missing HQMR data was only performed for determining the basis functions and these imputed values were not used in the rest of the model fitting process.

To determine the value of  $J$ , Lindstrom et al. (2011) calculate a number of regression statistics (MSE,  $R^2$ , AIC and BIC) via leave-one-column-out cross-validation for a range of values of  $J$ . Specifically, a column of the data matrix  $\mathbf{Y}$  is removed, the  $J$  basis functions are determined using the reduced data, then these  $J$  basis functions are used to predict the left-out column. This is repeated for each column of  $\mathbf{Y}$  to obtain the regression statistics for a given value of  $J$ . They then select the value of  $J$  that optimises the regression statistics. We used this approach for the HQMR data and the plots of the regression statistics for a range of values of  $J$  are displayed in Figure 6. These plots indicate that the appropriate choice is  $J = 1$  and the corresponding basis function is shown in Figure 7. However, a single basis function is unlikely to be able to capture all the temporal trends that may be present in the data. This is further evidenced by the partial autocorrelations of the residuals from regressions of the HQMR data on this single basis function, displayed in

---

**Algorithm 1:** Approach used by Szpiro et al. (2010) for deriving the temporal basis functions.

**initialize**

Normalise the columns of  $\mathbf{Y}$  to have mean 0 and variance 1;  
 Impute the missing values with the fitted values from a regression of each column of  $\mathbf{Y}$  on the single column vector given by the row averages of the non-missing values of  $\mathbf{Y}$ ;

**repeat**

Calculate the SVD of  $\mathbf{Y}$ ;  
 Update the imputed values with the fitted values from a regression of each column of  $\mathbf{Y}$  on the first  $J$  left singular vectors of the SVD;

**until** *The imputed values do not change;*

**smooth**

Use smoothing splines to smooth the first  $J$  left singular vectors from the SVD of the final converged  $\mathbf{Y}$ ;

---

168 Figure 8 for four randomly chosen stations. These partial autocorrelation plots clearly  
 169 show that there is some seasonal pattern still present in the residuals. Specifically, there  
 170 appears to be a distinct cyclic pattern over the “wet” and “dry” seasons.

171 To address these issues, we developed a new approach for deriving the temporal basis  
 172 functions that incorporates both a deterministic component and a data-derived component.  
 173 For the deterministic component we set the first three basis functions to be:

$$f_1(t) = \frac{2\pi t}{12}, \quad f_2(t) = \sin\left(\frac{2\pi t}{12}\right), \quad f_3(t) = \cos\left(\frac{2\pi t}{12}\right). \quad (7)$$

174 The form of these basis functions was chosen mainly to reflect the monthly nature of the  
 175 data and because these functions should explicitly capture seasonal patterns that cycle  
 176 over wet and dry seasons.

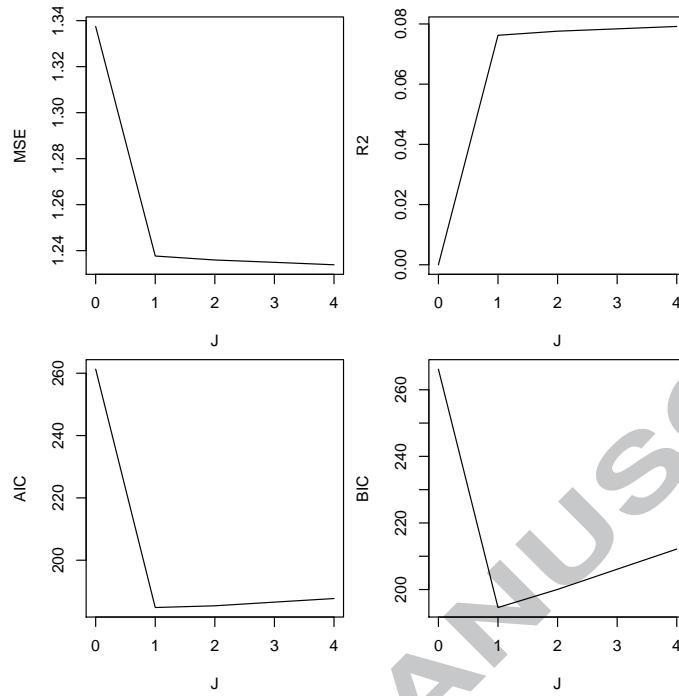


Figure 6: Cross-validated regression statistics (MSE,  $R^2$ , AIC and BIC) for a range of values of  $J$ .

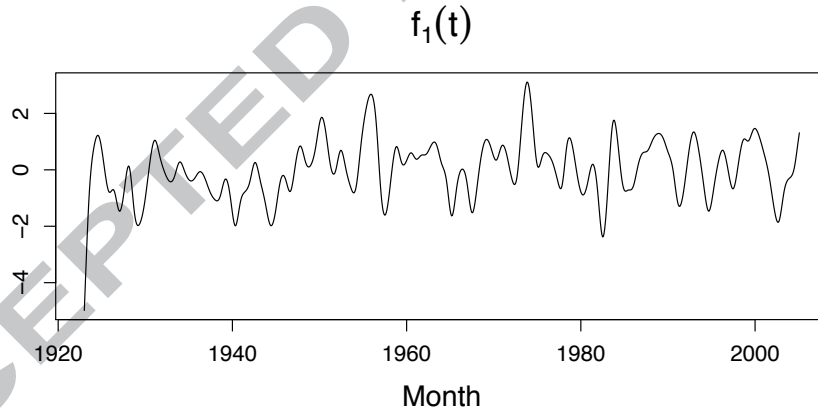


Figure 7: The basis function obtained by applying Algorithm 1 with  $J = 1$  to the HQMR data.

177 For the data-derived component, we applied Algorithm 1 to the residuals from a regres-  
 178 sion of the HQMR data on the three deterministic basis functions (including an intercept).

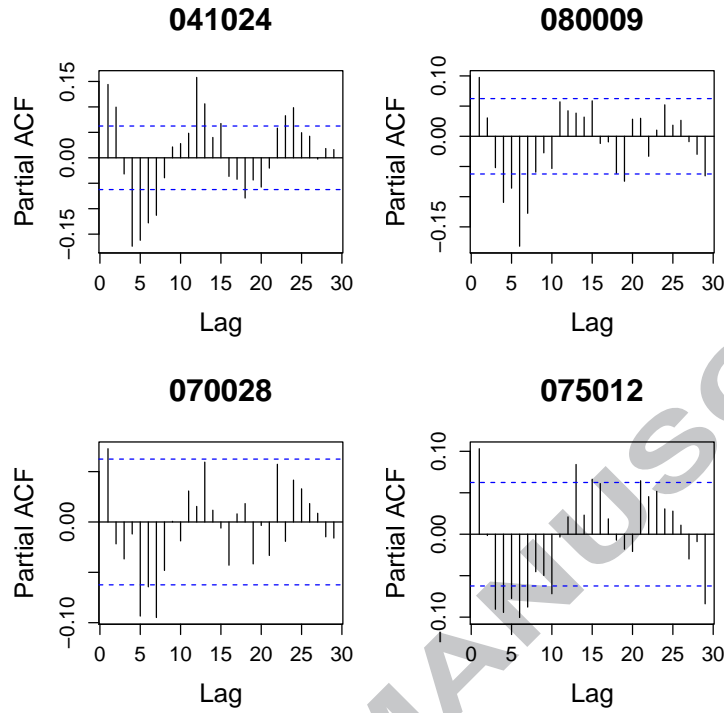


Figure 8: Partial autocorrelations of the residuals from a regression of the HQMR data on the basis function shown in Figure 7, for four randomly chosen stations.

Any basis functions found by the algorithm should capture other trends that still remain in the data. Based on the cross-validated regression statistics (not-shown), the algorithm chose one basis function for these residuals, which is displayed in Figure 9. This basis function is very similar to the single basis function found when applying Algorithm 1 to the HQMR data. This similarity shows that the data-derived basis function does not capture variability captured by the deterministic basis functions. The importance of including the deterministic basis functions in the model is confirmed below by comparing Figures 8 and 10 and by examining Figure 11.

Combining the deterministic and data-derived components, we chose four basis functions, namely, the three basis functions given in (7) and the basis function displayed in

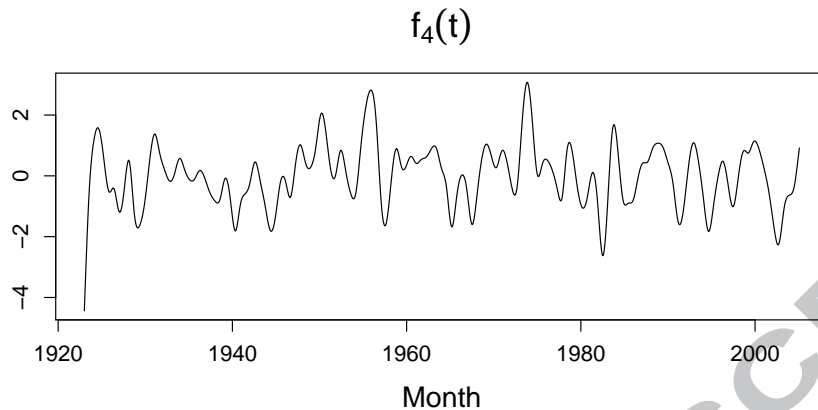


Figure 9: The basis function obtained by applying Algorithm 1 with  $J = 1$  to the residuals from a regression of the HQMR data on the three deterministic basis functions.

Figure 9. Displayed in Figure 10 are the partial autocorrelations of the residuals from regressions of the HQMR data on these four basis functions, for the four stations of Figure 8. Comparing these plots to the plots in Figure 8, we see that these four basis functions are much more effective in capturing the trends in the data, including the cyclic seasonal trend. We also investigated including an indicator for rainfall as a basis function, to help deal with the zero rainfall measurements. However, it did not lead to any improvements so we did not proceed with this further.

#### 4.2. Estimating the Parameters

The first step in estimating the model parameters is to estimate the coefficients for each basis function, i.e.,  $\beta_j = (\beta_j(s_1), \dots, \beta_j(s_N))^T$  for  $j = 0, \dots, J$ . Note that the coefficients for any given basis function will vary from station to station. These spatially-varying coefficients were estimated by regressing the HQMR data for each station on the four basis functions derived in Section 4.1. The  $-\log_{10}$ -transformed  $p$ -values for each coefficient are displayed in Figure 11. Note that these  $p$ -values obtained from the ordinary least squares regression are not proper  $p$ -values in the context of our hierarchical model given in (5)

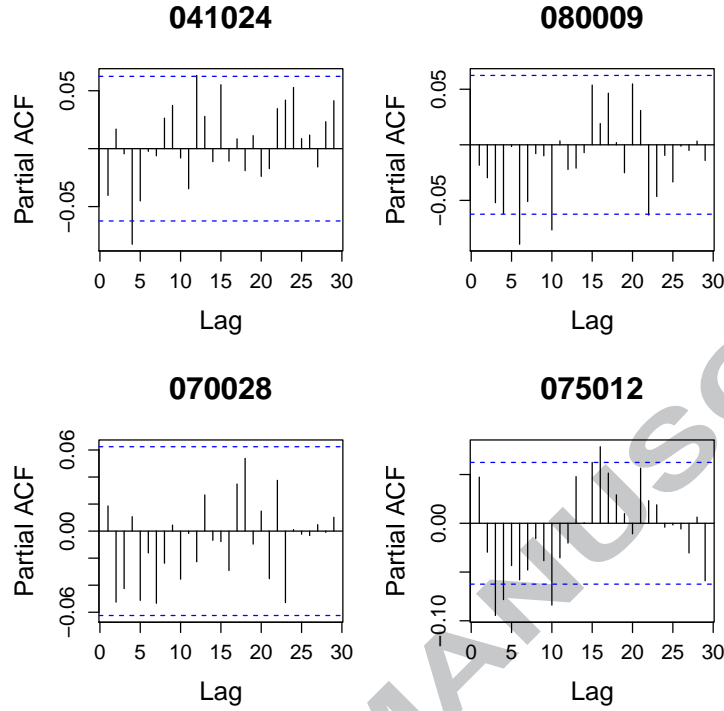


Figure 10: Partial autocorrelations of the residuals from a regression of the HQMR data on the combined deterministic and data-derived basis functions, for the four randomly chosen stations of Figure 8.

and (6), as they do not take into account the process of deriving the basis functions. Hence, they should be considered more as “indicative”  $p$ -values. The bubble-plots (which display the transformed  $p$ -values spatially) for  $\beta_2$ ,  $\beta_3$  and  $\beta_4$  are particularly interesting. These correspond to the spatial coefficients of the basis functions  $f_2(t) = \sin\left(\frac{2\pi t}{12}\right)$ ,  $f_3(t) = \cos\left(\frac{2\pi t}{12}\right)$  and  $f_4(t)$  given in Figure 9, respectively. Based on the spatial pattern of these  $p$ -values, in terms of capturing temporal trends present in the data it appears that  $f_2(t)$  is most significant for the southern region of stations,  $f_3(t)$  for the northern region of stations and  $f_4(t)$  for the central-eastern region of stations. This provides further evidence of the effectiveness of the chosen basis functions in modelling the spatial variability of the temporal trends present in the data. However, many of the  $p$ -values for  $\beta_1$ , which corresponds to



the basis function  $f_1(t) = \frac{2\pi t}{12}$ , are not significant. This may be an indication that this particular basis function is not needed in the model. At this stage, we will continue with the  $J = 4$  basis functions but will revisit this issue shortly.

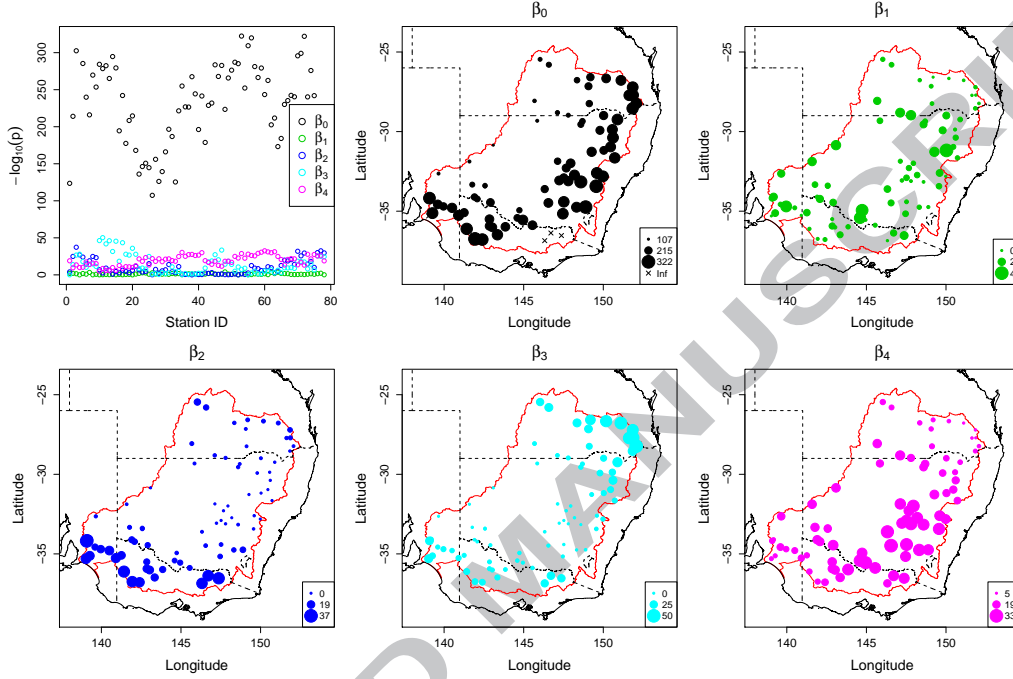


Figure 11: In the top-left panel, the  $-\log_{10}$ -transformed  $p$ -values for  $\beta_j(s_i)$  are plotted for all basis functions  $j = 0, \dots, J$  (with  $J = 4$ ) and all stations  $i = 1, \dots, N$ . In the remaining panels, the transformed  $p$ -values are plotted spatially as a bubble-plot for each basis function. Note that larger values of  $-\log_{10}(p)$  indicate greater significance.

The second step is to estimate the coefficients for each set of spatial covariates, i.e.,  $\alpha_j$  for  $j = 0, \dots, J$ . The model allows for the set of spatial covariates to differ for each value of  $j$ . However, for simplicity, we set the spatial covariates to be a station's Cartesian  $x$ - and  $y$ -coordinate, both measured in kilometres, elevation, measured in metres, the two indicators for the three climatic regimes, and the six interactions between these two indicators and the previous three variables, for all  $j$ . That is, for each  $j$ ,  $\mathbf{X}_j = \mathbf{X}$ , where  $\mathbf{X}$  is a  $78 \times 12$  matrix consisting of a column of 1's followed by columns of the  $x$ -coordinates,  $y$ -coordinates,

elevations, the two indicators and the six interactions for the 78 stations. The approach allows the set of covariates to be extended to include other environmental variables. These would need to be obtained from other data sets and linked to the HQMR data. We can estimate  $\alpha_j$  by regressing the estimates of  $\beta_j$  found in the first step on  $\mathbf{X}$ . The estimated  $\alpha_j$  for  $j = 0, \dots, J$ , along with bootstrap standard errors, are displayed in Table 1. Details on how the standard errors were calculated are given in Section 4.3. Based on these parameter estimates and their standard errors, there are a number of significant parameters across the values of  $j$ . However, the parameter estimates for  $j = 1$ ,  $\hat{\alpha}_1$ , are all extremely small with relatively large standard errors. This supports the previous observation that the basis function  $f_1(t) = \frac{2\pi t}{12}$  is not adding any further information to the model and may not be necessary.

The third step is to estimate the parameters of the covariance function for each  $\beta_j$ , i.e.,  $\theta_j$  for  $j = 0, \dots, J$ . This requires specifying an appropriate covariance function and the simplest approach for doing so is to construct empirical variograms. Empirical variograms were produced using the residuals from the regressions performed in the second step in estimating the  $\alpha_j$  and these are plotted in Figure 12. Directional empirical variograms constructed in both the north-south and east-west directions showed no evidence of anisotropy. Based on the plots in Figure 12, an exponential covariance function,  $\gamma(d) = \tau^2 + \sigma^2(1 - \exp(-\phi d))$  where  $d$  is the distance between stations, seems appropriate. The covariance parameters for each  $j$ ,  $\theta_j = (\tau_j^2, \sigma_j^2, \phi_j)^T$ , were estimated by weighted least squares using a range of initial values of  $\sigma^2$  and  $\phi$ . The estimates, along with bootstrap standard errors, are displayed in Table 2. We notice that there are no standard errors for  $j = 1$ , indicating that the same estimates were obtained for every bootstrap sample.

The fourth and last step is to estimate the parameters of the covariance function for the residual component  $\mathbf{e}_t = (e(s_1, t), \dots, e(s_N, t))^T$  given in (4), i.e.,  $\theta_e$ . Recall that one of the main assumptions of model (1) is that the temporal structure present in the data is captured by the  $\mu(s_i, t)$  term and that the residuals are therefore uncorrelated

	$j = 0$	$j = 1$	$j = 2$	$j = 3$	$j = 4$
Intercept ( $\hat{\alpha}_{j0}$ )	$2.87 \times 10^0$ ( $3.85 \times 10^{-1}$ )	$2.05 \times 10^{-4}$ ( $1.78 \times 10^{-4}$ )	$-7.91 \times 10^{-2}$ ( $2.41 \times 10^{-1}$ )	$2.70 \times 10^{-1}$ ( $3.47 \times 10^{-1}$ )	$2.03 \times 10^{-1}$ ( $4.30 \times 10^{-1}$ )
$x$ -coord ( $\hat{\alpha}_{j1}$ )	$5.72 \times 10^{-4}$ ( $1.04 \times 10^{-3}$ )	$8.24 \times 10^{-7}$ ( $4.17 \times 10^{-7}$ )	$2.98 \times 10^{-5}$ ( $6.22 \times 10^{-4}$ )	$2.11 \times 10^{-5}$ ( $8.77 \times 10^{-4}$ )	$3.15 \times 10^{-4}$ ( $2.04 \times 10^{-4}$ )
$y$ -coord ( $\hat{\alpha}_{j2}$ )	$-6.96 \times 10^{-4}$ ( $8.43 \times 10^{-4}$ )	$-4.62 \times 10^{-7}$ ( $3.46 \times 10^{-7}$ )	$5.98 \times 10^{-4}$ ( $5.21 \times 10^{-4}$ )	$1.05 \times 10^{-3}$ ( $7.32 \times 10^{-4}$ )	$-2.75 \times 10^{-4}$ ( $1.70 \times 10^{-4}$ )
Elevation ( $\hat{\alpha}_{j3}$ )	$9.68 \times 10^{-4}$ ( $5.25 \times 10^{-4}$ )	$-4.13 \times 10^{-7}$ ( $2.39 \times 10^{-7}$ )	$-1.05 \times 10^{-4}$ ( $2.91 \times 10^{-4}$ )	$-1.29 \times 10^{-4}$ ( $4.39 \times 10^{-4}$ )	$-4.06 \times 10^{-5}$ ( $1.15 \times 10^{-4}$ )
$I_1$ ( $\hat{\alpha}_{j4}$ )	$-4.27 \times 10^{-1}$ ( $4.38 \times 10^{-1}$ )	$6.90 \times 10^{-4}$ ( $1.76 \times 10^{-4}$ )	$8.24 \times 10^{-2}$ ( $2.70 \times 10^{-1}$ )	$-2.54 \times 10^{-1}$ ( $3.78 \times 10^{-1}$ )	$2.50 \times 10^{-1}$ ( $9.66 \times 10^{-2}$ )
$I_2$ ( $\hat{\alpha}_{j5}$ )	$-4.93 \times 10^{-1}$ ( $5.43 \times 10^{-1}$ )	$2.02 \times 10^{-4}$ ( $1.80 \times 10^{-4}$ )	$5.91 \times 10^{-2}$ ( $3.39 \times 10^{-1}$ )	$-1.77 \times 10^{-1}$ ( $4.75 \times 10^{-1}$ )	$2.17 \times 10^{-1}$ ( $1.14 \times 10^{-1}$ )
$I_1 \times x$ ( $\hat{\alpha}_{j6}$ )	$7.55 \times 10^{-4}$ ( $1.14 \times 10^{-3}$ )	$-1.20 \times 10^{-6}$ ( $4.72 \times 10^{-7}$ )	$9.27 \times 10^{-6}$ ( $6.82 \times 10^{-4}$ )	$2.97 \times 10^{-4}$ ( $9.96 \times 10^{-4}$ )	$-4.30 \times 10^{-4}$ ( $2.41 \times 10^{-4}$ )
$I_1 \times y$ ( $\hat{\alpha}_{j7}$ )	$4.70 \times 10^{-4}$ ( $8.89 \times 10^{-4}$ )	$2.55 \times 10^{-7}$ ( $3.79 \times 10^{-7}$ )	$-1.18 \times 10^{-4}$ ( $5.61 \times 10^{-4}$ )	$-3.72 \times 10^{-4}$ ( $7.82 \times 10^{-4}$ )	$1.69 \times 10^{-4}$ ( $1.82 \times 10^{-4}$ )
$I_1 \times \text{Elev}$ ( $\hat{\alpha}_{j8}$ )	$1.58 \times 10^{-5}$ ( $6.63 \times 10^{-4}$ )	$-7.08 \times 10^{-7}$ ( $3.12 \times 10^{-7}$ )	$-5.24 \times 10^{-5}$ ( $3.25 \times 10^{-4}$ )	$3.09 \times 10^{-4}$ ( $5.17 \times 10^{-4}$ )	$-1.86 \times 10^{-4}$ ( $1.33 \times 10^{-4}$ )
$I_2 \times x$ ( $\hat{\alpha}_{j9}$ )	$-9.34 \times 10^{-5}$ ( $1.12 \times 10^{-3}$ )	$-7.63 \times 10^{-7}$ ( $4.48 \times 10^{-7}$ )	$2.32 \times 10^{-4}$ ( $6.51 \times 10^{-4}$ )	$2.21 \times 10^{-4}$ ( $9.23 \times 10^{-4}$ )	$-1.22 \times 10^{-4}$ ( $2.15 \times 10^{-4}$ )
$I_2 \times y$ ( $\hat{\alpha}_{j10}$ )	$-2.71 \times 10^{-4}$ ( $1.34 \times 10^{-3}$ )	$8.37 \times 10^{-7}$ ( $4.28 \times 10^{-7}$ )	$-6.84 \times 10^{-5}$ ( $7.79 \times 10^{-4}$ )	$-3.93 \times 10^{-4}$ ( $1.12 \times 10^{-3}$ )	$3.71 \times 10^{-4}$ ( $2.57 \times 10^{-4}$ )
$I_2 \times \text{Elev}$ ( $\hat{\alpha}_{j11}$ )	$4.77 \times 10^{-4}$ ( $9.95 \times 10^{-4}$ )	$6.82 \times 10^{-7}$ ( $4.03 \times 10^{-7}$ )	$-1.66 \times 10^{-4}$ ( $5.89 \times 10^{-4}$ )	$-7.01 \times 10^{-6}$ ( $8.51 \times 10^{-4}$ )	$-1.14 \times 10^{-4}$ ( $2.01 \times 10^{-4}$ )

Table 1: Estimates of  $\alpha_j$  for  $j = 0, \dots, J$  (with  $J = 4$ ). Standard errors are given in brackets.

over time. That is, if there were no missing values in the data, each  $\mathbf{e}_t$  would have a common covariance matrix  $\Sigma_e(\theta_e)$ . Therefore, each  $\mathbf{e}_t$  can be treated as an independent realisation for estimating the covariance parameters  $\theta_e$ . Again, we first need to specify an

	$j = 0$	$j = 1$	$j = 2$	$j = 3$	$j = 4$
$\hat{\tau}_j^2$	$2.49 \times 10^{-2}$ ( $5.03 \times 10^{-2}$ )	0 (-)	0 ( $9.11 \times 10^{-3}$ )	0 ( $3.84 \times 10^{-3}$ )	0 ( $1.71 \times 10^{-3}$ )
$\hat{\sigma}_j^2$	$1.19 \times 10^{-2}$ ( $3.85 \times 10^{-1}$ )	$7.00 \times 10^{-8}$ (-)	$3.43 \times 10^{-3}$ ( $1.50 \times 10^{-2}$ )	$4.18 \times 10^{-3}$ ( $2.85 \times 10^{-2}$ )	$1.28 \times 10^{-3}$ ( $1.28 \times 10^{-3}$ )
$\hat{\phi}_j$	250 ( $7.57 \times 10^3$ )	250 (-)	171 (16.5)	81.5 (22.9)	250 ( $5.02 \times 10^{-7}$ )

Table 2: Estimates of  $\theta_j$  for  $j = 0, \dots, J$  (with  $J = 4$ ). Standard errors are given in brackets.

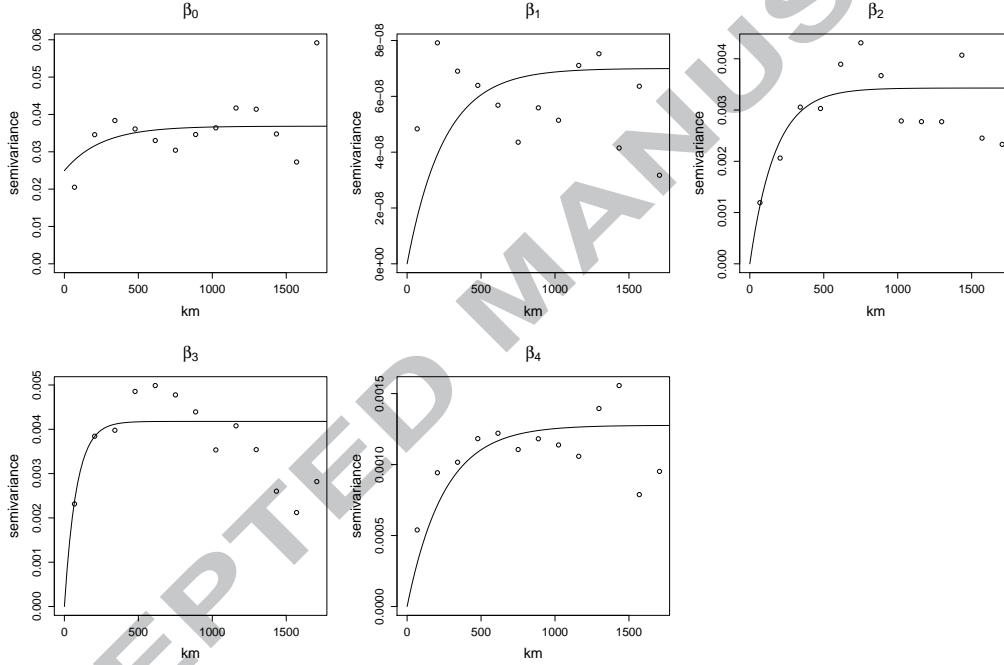


Figure 12: Empirical variograms of the residuals from regressing  $\hat{\beta}_j$  on  $\mathbf{X}$ , for  $j = 0, \dots, J$  (with  $J = 4$ ). Superimposed onto each plot is a fitted exponential variogram.

appropriate covariance function. Empirical variograms were calculated for the residuals from the regressions performed in the first step to estimate the  $\beta_j$ . Displayed in Figure 13 are the variograms for four randomly selected time points. These plots indicate that

an exponential covariance function may also be appropriate for the residual component. Directional empirical variograms (north-south and east-west directions) of the residuals averaged over time showed no obvious evidence of anisotropy. The residual covariance parameters,  $\theta_e = (\tau_e^2, \sigma_e^2, \phi_e)$ , were estimated via maximum likelihood using the `likfit` function in the R package `geoR`, as this allowed us to utilise all the residuals,  $e_t$  for  $t = 1, \dots, T$ , in the estimation. Using multiple initial values for  $\sigma_e^2$  and  $\phi_e$ , the estimates and their bootstrap standard errors are displayed in Table 3.

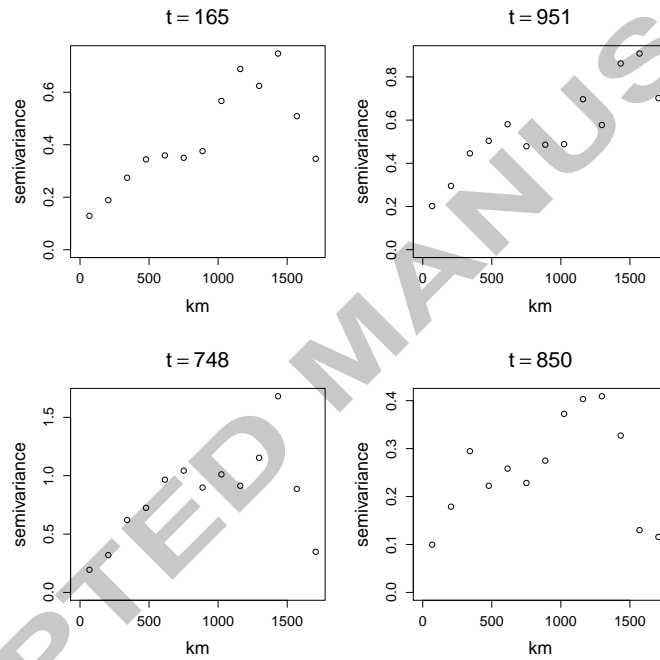


Figure 13: Empirical variograms of the residuals from regressing the HQMR data on the four basis functions, for four randomly selected time points.

Having obtained the parameter estimates for our model, it is worthwhile checking whether the model assumptions hold. We can see directly from (6) that our model assumes both  $\beta_j$  and  $e_t$  follow a normal distribution. Displayed in Figure 14 are the normal Q-Q plots of the standardised residuals obtained from the regression of  $\beta_j$  on  $\mathbf{X}$ , for

$\hat{\tau}_e^2$	$\hat{\sigma}_e^2$	$\hat{\phi}_e$
0.0751	0.756	417
(0.0411)	(0.0593)	(48.1)

Table 3: Estimates of  $\theta_e$  for  $J = 4$  basis functions. Standard errors are given in brackets.

268  $j = 0, \dots, 4$ . These plots show that the  $\beta_j$  are approximately normally distributed. In  
 269 Figure 15, a normal Q-Q plot of  $e_t$  (after standardising) for all  $t$  is displayed in the left  
 270 panel. Displayed in the right panel is a Q-Q plot of the residuals against a  $t$ -distribution  
 271 with 6 degrees of freedom. These plots indicate that the distribution of the  $e_t$ , although  
 272 longer-tailed than a normal distribution, is still symmetric and is well approximated by  
 273 a  $t$ -distribution. While further research into how to incorporate this kind of distribution  
 274 into the analysis is needed, its impact on the present analysis is likely to be some loss of  
 275 efficiency in the parameter estimates (without undermining their validity).

#### 276 4.3. Standard Error Estimation

277 We used an approach that involved bootstrapping the HQMR data to estimate the  
 278 standard errors of the parameter estimates  $\hat{\alpha}_j$  and  $\hat{\theta}_j$ , for  $j = 0, \dots, J$ , and  $\hat{\theta}_e$ . Due to the  
 279 spatio-temporal nature of our data, generating a bootstrap sample required bootstrapping  
 280 in both the spatial and temporal dimensions. Our approach was designed to ensure that the  
 281 bootstrap samples introduced adequate variability while maintaining spatial and temporal  
 282 relationships that may be present in the original data. We used a non-parametric bootstrap  
 283 to avoid making any assumptions on the distributional shape of the data.

284 Bootstrapping in the temporal dimension requires selecting a bootstrap sample of  $T$   
 285 months, denoted as  $M_b$ , from the original set of months,  $M = \{1, \dots, T\}$ . In order to  
 286 maintain any temporal patterns present in the data, we first defined a “temporal selection  
 287 block” as a block of 5 contiguous years or 60 months. The second step was to divide the  
 288 original set of months  $M$  into a set of overlapping blocks by sliding the temporal selection

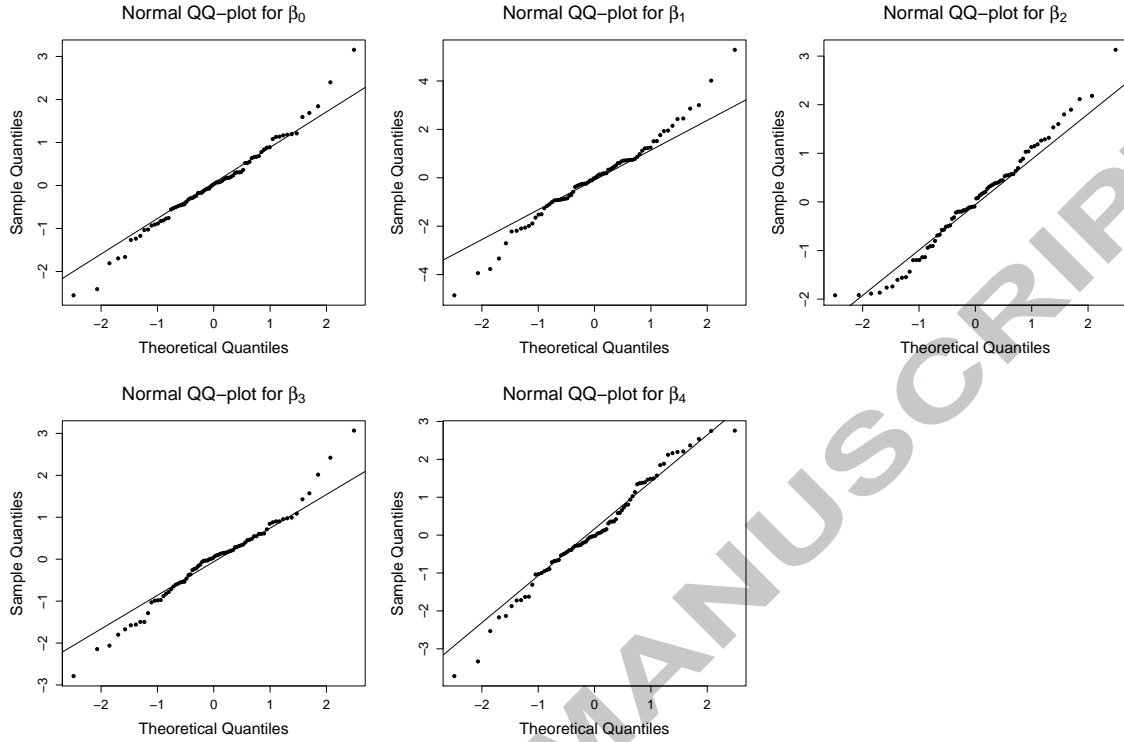


Figure 14: Normal Q-Q plot of the (standardised) residuals from regressing  $\beta_j$  on  $\mathbf{X}$ , for  $j = 0, \dots, 4$ .

289 block along the months by 1 year increments. For our data with  $T = 986$ , this produced  
 290 78 possible temporal selection blocks. Next we randomly selected blocks from these 78  
 291 possible blocks, with replacement, until the total number of months in the selected blocks  
 292 was at least  $T$ . Finally, the bootstrap sample  $M_b$  was generating by concatenating the  
 293 selected blocks end-to-end. We note that there is a trade-off in choosing the block size.  
 294 That is, larger block sizes will increase the chance of maintaining temporal patterns but  
 295 will decrease the variability in the bootstrap samples, and vice versa for smaller block sizes.  
 296 A similar approach was used to bootstrap in the spatial dimension. However, some  
 297 extra care is required due to the spatial covariates that are associated with each station.  
 298 The first step was to define a set of overlapping spatial blocks by moving an approximately  
 299 rectangular geographic region over the  $N = 78$  stations in the MDB. The blocks were

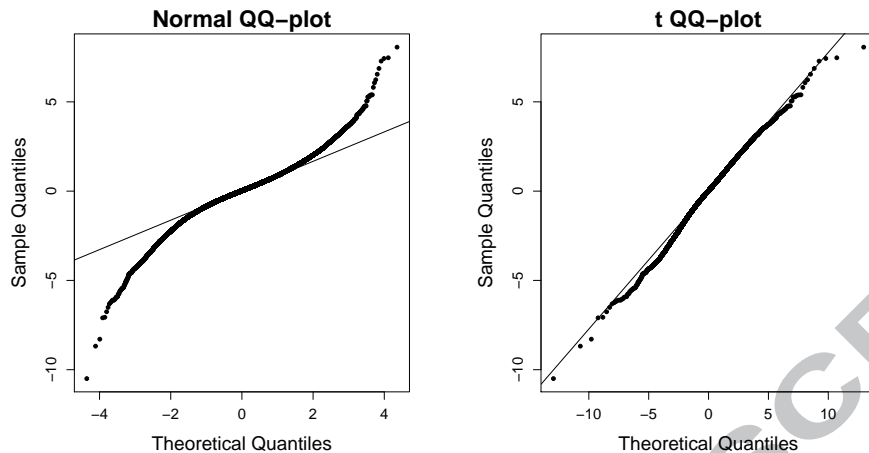


Figure 15: Q-Q plot of (standardised)  $e_t$ , over all  $t$ , against a normal distribution (left) and a  $t$ -distribution with 6 degrees of freedom (right).

defined so that each contained approximately 10 stations, with any two overlapping blocks sharing at least a few stations. A total of 22 such overlapping blocks were obtained and these are displayed in the left panel of Figure 16. These blocks will serve as the sampling unit for generating a bootstrap sample of stations,  $S_b$ . The second step was to divide the  $N = 78$  stations into a set of 8 non-overlapping spatial blocks, again each with approximately 10 stations. These non-overlapping blocks are displayed in the right panel of Figure 16. Finally, to generate the bootstrap sample  $S_b$ , we follow the process described in Algorithm 2 below:

To produce all the standard errors displayed in Tables 1 to 3, we generated 200 bootstrap samples of HQMR data and fitted the model on each sample to obtain the parameter estimates. Bootstrapping the HQMR data in this manner is not the only way to obtain standard error estimates. For example, we could bootstrap the residuals or we could use a parametric bootstrap approach. However, we prefer our approach as it does not require making any normality assumptions on the residuals or any underlying parametric model assumptions. This is further supported by the fact that the residuals  $e_t$  seemed to have a longer tail than a normal distribution, as shown in Figure 15.



---

**Algorithm 2:** Selecting a bootstrap sample of stations.

---

**repeat**

Select a non-overlapping block from Figure 16;  
 Randomly select with replacement an overlapping block from Figure 16;  
 Replace the HQMR data for the stations in the non-overlapping block with the  
 HQMR data for stations in the overlapping block;  
 If the overlapping block contains fewer stations than the non-overlapping block,  
 recycle through the stations in the overlapping block;

**until** *All non-overlapping blocks have been selected;*

---

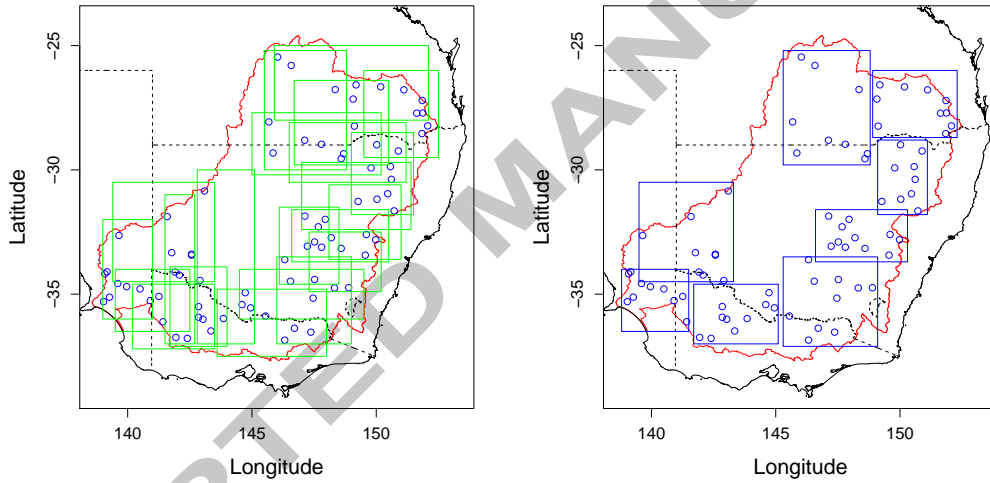


Figure 16: Overlapping (left) and non-overlapping (right) spatial blocks used for generating the bootstrap sample of stations.

#### 4.4. Final Estimated Model

The  $p$ -values for  $\beta_1$  displayed in Figure 11 and the parameter estimates for  $\alpha_1$  given in Table 1 provide evidence that the basis function  $f_1(t) = \frac{2\pi t}{12}$  is not improving the model fit. Therefore, we removed this basis function and our final model, which we use for the remainder of this paper, consists of  $J = 3$  basis functions. That is, the following two basis

functions:

$$f_1(t) = \sin\left(\frac{2\pi t}{12}\right), \quad f_2(t) = \cos\left(\frac{2\pi t}{12}\right), \quad (8)$$

and the corresponding data-derived basis function, displayed in Figure 17. The  $p$ -values for the  $\beta_j$  for  $j = 0, \dots, 3$  are displayed in Figure 18 and the remaining parameter estimates are given in Tables 4 to 6.

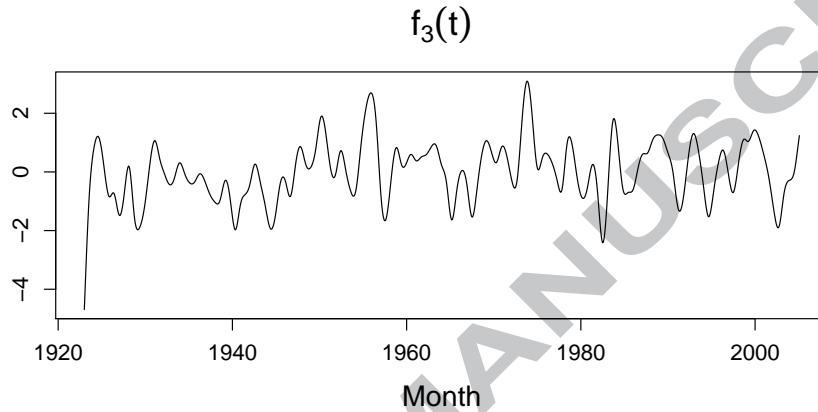


Figure 17: The basis function obtained by applying Algorithm 1 with  $J = 1$  to the residuals from a regression of the HQMR data on the two deterministic basis functions given in (8).

## 5. Prediction

There are generally two types of predictions that are of interest. The first is temporal prediction, where the goal is to make predictions for future time points, usually at locations where past data have been observed. The second is spatial prediction, where the goal is to make predictions at locations where no past data have been observed, usually for time points where data from other locations are available. Our model allows both types of predictions in a natural way. Note that the predictions we produce are at the cube-root transformed scale. Back-transformed predictions on the original scale can be obtained by cubing the predicted values. To preserve the mean of the predicted values on the

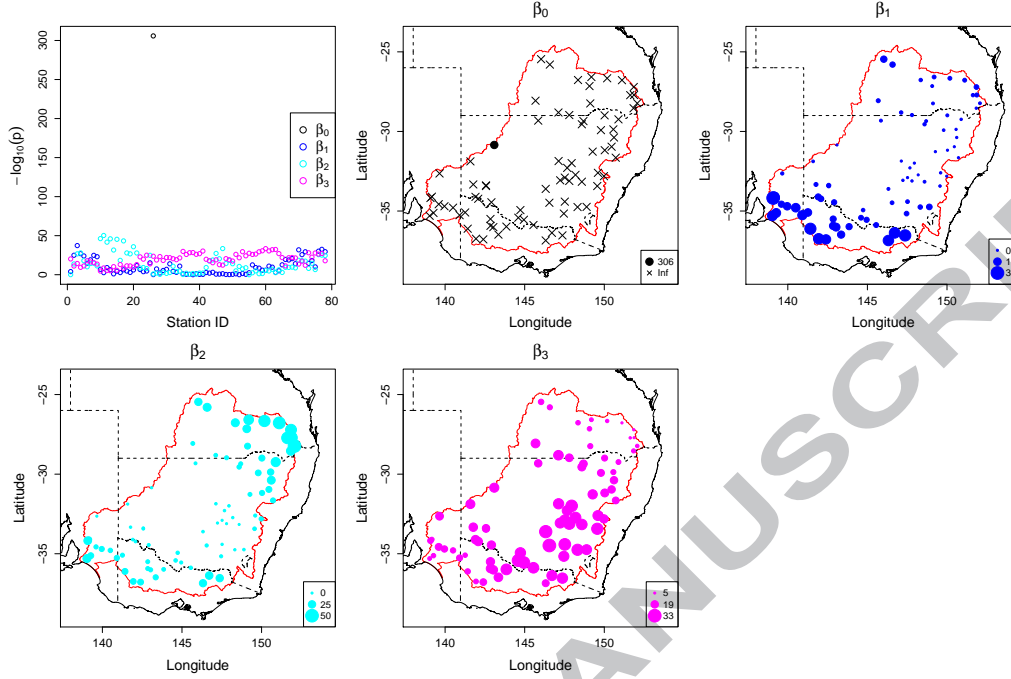


Figure 18: In the top-left panel, the  $-\log_{10}$ -transformed  $p$ -values for  $\beta_j(\mathbf{s}_i)$  are plotted for all basis functions  $j = 0, \dots, J$  (with  $J = 3$ ) and all stations  $i = 1, \dots, N$ . In the remaining panels, the transformed  $p$ -values are plotted spatially as a bubble-plot for each basis function. Again larger values of  $-\log_{10}(p)$  indicate greater significance.

original scale, adjusted back-transformations such as  $\hat{Y}^3 + 3\hat{Y}\hat{\sigma}^2$ , where  $\hat{Y}$  and  $\hat{\sigma}^2$  denote the prediction and the estimated variance of  $Y$  on the transformed scale, or the smearing estimator of Duan (1983) can be used.

### 5.1. Temporal Prediction

As the HQMR data we are analysing are monthly, we will describe temporal prediction in this context. However, the procedure can be generalised to data measured at other frequencies, e.g., weekly or daily. Let  $\mathbf{s}_o$  denote a particular station for which past data have been observed and let  $t_u$  denote the index of an unobserved future time point. Given parameter estimates  $\hat{\beta}_j$  for  $j = 0, \dots, 3$ , we can predict the value of  $Y(\mathbf{s}_o, t_u)$  using equation

	$j = 0$	$j = 1$	$j = 2$	$j = 3$
Intercept ( $\hat{\alpha}_{j0}$ )	$2.93 \times 10^0$ ( $3.79 \times 10^{-1}$ )	$-7.91 \times 10^{-2}$ ( $2.41 \times 10^{-1}$ )	$2.70 \times 10^{-1}$ ( $3.47 \times 10^{-1}$ )	$2.04 \times 10^{-1}$ ( $4.31 \times 10^{-1}$ )
$x$ -coordinate ( $\hat{\alpha}_{j1}$ )	$7.86 \times 10^{-4}$ ( $1.04 \times 10^{-3}$ )	$2.93 \times 10^{-5}$ ( $6.22 \times 10^{-4}$ )	$2.17 \times 10^{-5}$ ( $8.77 \times 10^{-4}$ )	$3.34 \times 10^{-4}$ ( $2.04 \times 10^{-4}$ )
$y$ -coordinate ( $\hat{\alpha}_{j2}$ )	$-8.15 \times 10^{-4}$ ( $8.52 \times 10^{-4}$ )	$5.98 \times 10^{-4}$ ( $5.21 \times 10^{-4}$ )	$1.05 \times 10^{-3}$ ( $7.32 \times 10^{-4}$ )	$-2.83 \times 10^{-4}$ ( $1.70 \times 10^{-4}$ )
Elevation ( $\hat{\alpha}_{j3}$ )	$8.62 \times 10^{-4}$ ( $5.22 \times 10^{-4}$ )	$-1.05 \times 10^{-4}$ ( $2.91 \times 10^{-4}$ )	$-1.29 \times 10^{-4}$ ( $4.39 \times 10^{-4}$ )	$-5.22 \times 10^{-5}$ ( $1.15 \times 10^{-4}$ )
$I_1$ ( $\hat{\alpha}_{j4}$ )	$-2.48 \times 10^{-1}$ ( $4.35 \times 10^{-1}$ )	$8.23 \times 10^{-2}$ ( $2.70 \times 10^{-1}$ )	$-2.53 \times 10^{-1}$ ( $3.78 \times 10^{-1}$ )	$2.65 \times 10^{-1}$ ( $9.71 \times 10^{-2}$ )
$I_2$ ( $\hat{\alpha}_{j5}$ )	$-4.40 \times 10^{-1}$ ( $5.40 \times 10^{-1}$ )	$5.93 \times 10^{-2}$ ( $3.39 \times 10^{-1}$ )	$-1.77 \times 10^{-1}$ ( $4.75 \times 10^{-1}$ )	$2.19 \times 10^{-1}$ ( $1.14 \times 10^{-1}$ )
$I_1 \times x$ ( $\hat{\alpha}_{j6}$ )	$4.44 \times 10^{-4}$ ( $1.15 \times 10^{-3}$ )	$9.96 \times 10^{-6}$ ( $6.82 \times 10^{-4}$ )	$2.96 \times 10^{-4}$ ( $9.96 \times 10^{-4}$ )	$-4.58 \times 10^{-4}$ ( $2.42 \times 10^{-4}$ )
$I_1 \times y$ ( $\hat{\alpha}_{j7}$ )	$5.36 \times 10^{-4}$ ( $8.93 \times 10^{-4}$ )	$-1.18 \times 10^{-4}$ ( $5.61 \times 10^{-4}$ )	$-3.72 \times 10^{-4}$ ( $7.82 \times 10^{-4}$ )	$1.73 \times 10^{-4}$ ( $1.83 \times 10^{-4}$ )
$I_1 \times \text{Elev}$ ( $\hat{\alpha}_{j8}$ )	$-1.68 \times 10^{-4}$ ( $6.56 \times 10^{-4}$ )	$-5.25 \times 10^{-5}$ ( $3.25 \times 10^{-4}$ )	$3.08 \times 10^{-4}$ ( $5.17 \times 10^{-4}$ )	$-2.04 \times 10^{-4}$ ( $1.33 \times 10^{-4}$ )
$I_2 \times x$ ( $\hat{\alpha}_{j9}$ )	$-2.90 \times 10^{-4}$ ( $1.12 \times 10^{-3}$ )	$2.33 \times 10^{-4}$ ( $6.51 \times 10^{-4}$ )	$2.20 \times 10^{-4}$ ( $9.23 \times 10^{-4}$ )	$-1.44 \times 10^{-4}$ ( $2.16 \times 10^{-4}$ )
$I_2 \times y$ ( $\hat{\alpha}_{j10}$ )	$-5.55 \times 10^{-5}$ ( $1.34 \times 10^{-3}$ )	$-6.89 \times 10^{-5}$ ( $7.79 \times 10^{-4}$ )	$-3.92 \times 10^{-4}$ ( $1.12 \times 10^{-3}$ )	$3.89 \times 10^{-4}$ ( $2.57 \times 10^{-4}$ )
$I_2 \times \text{Elev}$ ( $\hat{\alpha}_{j11}$ )	$6.52 \times 10^{-4}$ ( $9.89 \times 10^{-4}$ )	$-1.66 \times 10^{-4}$ ( $5.89 \times 10^{-4}$ )	$-5.79 \times 10^{-6}$ ( $8.51 \times 10^{-4}$ )	$-9.02 \times 10^{-5}$ ( $2.01 \times 10^{-4}$ )

Table 4: Estimates of  $\alpha_j$  for  $j = 0, \dots, J$  (with  $J = 3$ ). Standard errors are given in brackets.

343 (5) as follows:

$$\hat{Y}(\mathbf{s}_o, t_u) = \hat{\beta}_0(\mathbf{s}_o) + \hat{\beta}_1(\mathbf{s}_o)f_1(t_u) + \hat{\beta}_2(\mathbf{s}_o)f_2(t_u) + \hat{\beta}_3(\mathbf{s}_o)f_3(t_u). \quad (9)$$

	$j = 0$	$j = 1$	$j = 2$	$j = 3$
$\hat{\tau}_j^2$	$1.52 \times 10^{-2}$ ( $4.81 \times 10^{-2}$ )	0 ( $9.12 \times 10^{-3}$ )	0 ( $3.88 \times 10^{-3}$ )	0 ( $1.39 \times 10^{-3}$ )
$\hat{\sigma}_j^2$	$1.48 \times 10^{-2}$ ( $5.11 \times 10^{-2}$ )	$3.42 \times 10^{-3}$ ( $1.50 \times 10^{-2}$ )	$4.18 \times 10^{-3}$ ( $2.85 \times 10^{-2}$ )	$1.34 \times 10^{-3}$ ( $1.92 \times 10^{-3}$ )
$\hat{\phi}_j$	250 (48.4)	171 (16.5)	81.3 (22.7)	123 (15.7)

Table 5: Estimates of  $\theta_j$  for  $j = 0, \dots, J$  (with  $J = 3$ ). Standard errors are given in brackets.

$\hat{\tau}_e^2$	$\hat{\sigma}_e^2$	$\hat{\phi}_e$
0.0761 (0.0412)	0.758 (0.0593)	418 (48.2)

Table 6: Estimates of  $\theta_e$  for  $J = 3$  basis functions. Standard errors are given in brackets.

The values of the deterministic basis functions,  $f_1(t_u)$  and  $f_2(t_u)$ , can be calculated by substituting  $t_u$  into the appropriate function given in (8). Hence only the value of the data-derived basis function,  $f_3(t_u)$ , needs to be determined in order to obtain the predicted value  $\hat{Y}(\mathbf{s}_o, t_u)$ .

Due to the seasonal nature of rainfall, there will generally be some correlations between rainfall values in the same month across different years. Consequently, our approach for determining  $f_3(t_u)$  is based on pooling similar information across years. Specifically, let  $\Upsilon$  denote a specified set of years (for which HQMR data have been observed) and let  $M_\Upsilon$  denote the set of indices for the same month corresponding to the time point  $t_u$  that belong in  $\Upsilon$ . For example, if  $t_u$  corresponded to January 2016 and  $\Upsilon$  was the set of years from 2011 to 2015, then  $M_\Upsilon$  would correspond to all January months in this five year span. We

355 then set the value of  $f_3(t_u)$  to be

$$f_3(t_u) = \frac{\sum_{t \in M_{\Upsilon}} f_3(t)}{|M_{\Upsilon}|}. \quad (10)$$

356 Clearly, how  $\Upsilon$  (and therefore  $M_{\Upsilon}$ ) is chosen will affect the value of  $f_3(t_u)$ . There is a  
 357 balance between choosing  $\Upsilon$  to be locally or globally focused, resulting in a bias/variance  
 358 tradeoff. On one hand, selecting  $\Upsilon$  to consist of only a small number of years close to  $t_u$   
 359 may be beneficial due to local correlations that may be present in the data. On the other  
 360 hand, selecting  $\Upsilon$  to consist of a larger number of years may be preferable as we would be  
 361 utilising more of the available data.

362 To evaluate the temporal predictive performance of our model, we set aside the last  
 363 twelve months of the HQMR data as a test set. Using the remaining data as the training  
 364 set, the model given in (5) was fitted to produce estimates  $\hat{\beta}_j$  for  $j = 0, \dots, 3$ . Using these  
 365 estimates, predicted values for each observation in the test set were calculated according  
 366 to (9) and (10). Four different specifications for  $\Upsilon$  were used, namely, the most recent year  
 367 in the training set, the most recent ten years in the training set, ten randomly selected  
 368 years from the training set, and all years in the training set. These predicted values are  
 369 displayed in Figure 19 for four randomly selected stations in the test set.

370 Based on the plots in Figure 19, we see that overall, the predicted values are relatively  
 371 close to the true observed values. The root-mean-square error (RMSE) across all obser-  
 372 vations in the test set was 0.945 (when setting  $\Upsilon$  to be ten randomly selected years). We  
 373 note that due to the smooth nature of the model, there are some discrepancies between the  
 374 predicted and true observed values at extreme points. However, any predictive model will  
 375 struggle to accurately predict extreme outlier values. It is also evident from these plots  
 376 that the predicted values are quite similar for all four specifications for  $\Upsilon$ . This indicates  
 377 that the deterministic seasonal trends are likely the dominant feature of the data, at least  
 378 for predictive purposes. Therefore, for these HQMR data, choosing  $\Upsilon$  to consist of 5-10  
 379 years (the most recent or a random selection) is likely to be sufficient. As a comparison,  
 380 using a moving average over the previous six months to predict the rainfall for each month

in the test set produced an RMSE of 1.251.

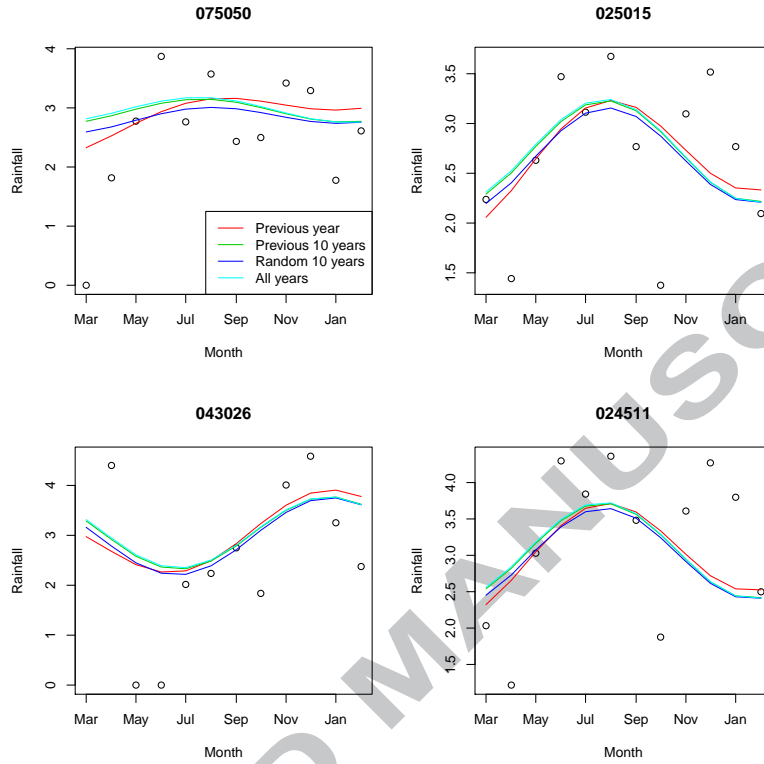


Figure 19: Predicted values for four randomly selected stations in the test set. Predicted values were calculated using equations (9) and (10) with four different specified sets for  $\Upsilon$ . The observed HQMR values are also displayed for each station.

## 5.2. Spatial Prediction

For spatial prediction, let  $s_u$  denote a station where no past data have been observed but for which the values of the spatial covariates, i.e., the station's Cartesian  $x$ - and  $y$ -coordinates, elevation, the two indicators for the three climatic regimes, and the six interactions are known. Also, let  $t_o$  denote the index of some previous time point for which data from other stations have been observed. Given that  $t_o$  corresponds to a time point where observed data exist at other stations, the values of all basis functions,  $f_1(t_o)$ ,  $f_2(t_o)$

and  $f_3(t_o)$ , are known. Therefore, if the parameter estimates for these basis functions, namely,  $\hat{\beta}_0(\mathbf{s}_u)$ ,  $\hat{\beta}_1(\mathbf{s}_u)$ ,  $\hat{\beta}_2(\mathbf{s}_u)$  and  $\hat{\beta}_3(\mathbf{s}_u)$ , were also known, then we could use equation (9) to predict the value of  $Y(\mathbf{s}_u, t_o)$ :

$$\hat{Y}(\mathbf{s}_u, t_o) = \hat{\beta}_0(\mathbf{s}_u) + \hat{\beta}_1(\mathbf{s}_u)f_1(t_o) + \hat{\beta}_2(\mathbf{s}_u)f_2(t_o) + \hat{\beta}_3(\mathbf{s}_u)f_3(t_o). \quad (11)$$

Given parameter estimates for the spatial covariates, i.e.,  $\hat{\boldsymbol{\alpha}}_j = (\hat{\alpha}_{j0}, \dots, \hat{\alpha}_{j11})^T$  for  $j = 0, \dots, J$ , we can determine the parameter estimates for the basis functions from equation (3). That is, letting  $X_{1u}, \dots, X_{11u}$  denote the unobserved station's covariate values, the parameter estimate  $\hat{\beta}_j(\mathbf{s}_u)$  can be calculated as

$$\hat{\beta}_j(\mathbf{s}_u) = \hat{\alpha}_{j0} + \sum_{k=1}^{11} \hat{\alpha}_{jk} X_{ku}. \quad (12)$$

To demonstrate how well our model performs spatial prediction, leave-one-out cross-validation was used to predict values for every station. Specifically, for each left-out station, our model was fitted using the remaining 77 stations to produce the basis functions and the estimates  $\hat{\boldsymbol{\alpha}}_j$  for  $j = 0, \dots, 3$ . From these estimates, the parameter estimates for the basis functions for the left-out station were determined according to (12). Subsequently, predicted values for every observation in the left-out station were calculated using (11). Predicted values for a randomly chosen station are displayed in Figure 20. From these plots we see that our model is able to capture the overall trend for this unobserved station quite well, with the largest differences between observed and predicted values occurring at the extreme outliers. The predicted values and observed values, both averaged over time, for all stations are displayed spatially in Figure 21. The predicted and observed values match up well, with the largest discrepancies again occurring for stations where the averaged values are very high or very low. The RMSE for all observations was 1.08. We also used observations in the nearest station as a naive predictor for each left-out station and this produced an RMSE of 0.65. The naive predictor having a lower RMSE is mainly due to the homogeneity in rainfall patterns and measurements in closely neighbouring stations.



412 However, a key advantage that our model has over any nearest neighbour-based approach  
 413 is that our model can also produce predictions at a future time point for a new spatial  
 414 location with no past data.

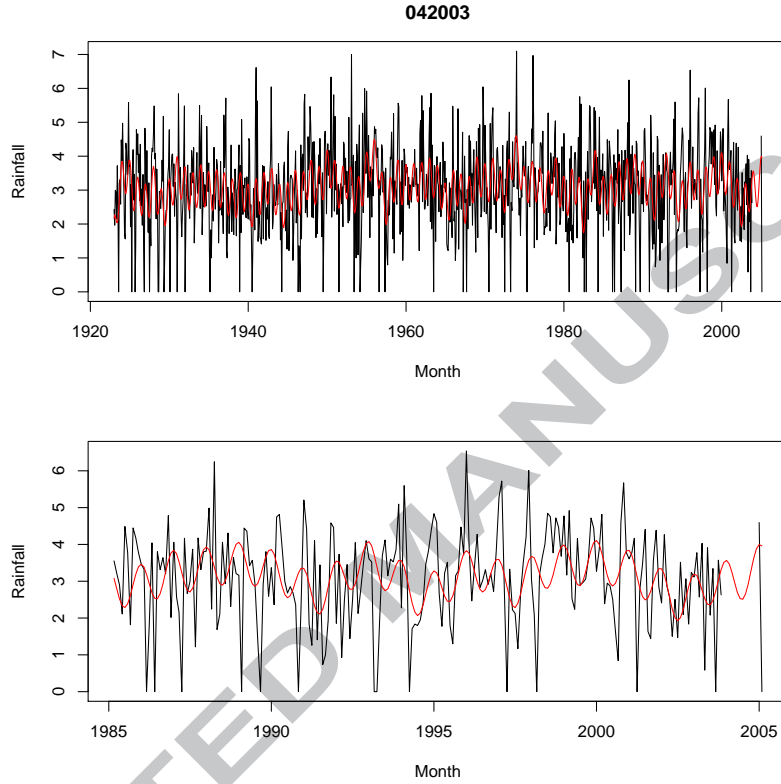


Figure 20: Predicted values, calculated using equations (11) and (12), for the test set station. The observed HQMR values for the station are displayed in black. The top plot displays values for the entire time range of the data and the bottom plot zooms in to the most recent 20 years.

## 415 6. Conclusions

416 In this paper, we proposed a non-Bayesian hierarchical model for analysing spatio-  
 417 temporal monthly rainfall data in the Murray-Darling Basin. Our methodology, based on  
 418 an approach proposed by Szpiro et al. (2010), models the monthly rainfall measurements

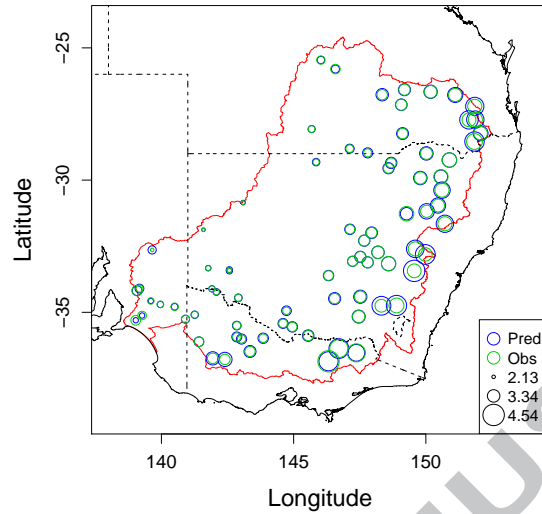


Figure 21: Predicted (blue) and observed (green) values, averaged over time, displayed spatially as a bubble plot.

419 observed at a particular spatial location as a linear combination of a set of basis functions.  
 420 Each basis function can be thought of as a particular temporal pattern that is shared  
 421 across a set of spatial locations. By setting the basis functions to be a novel combination  
 422 of both deterministic and data-derived functions we were able to capture much of the  
 423 temporal structure present in the rainfall data. Spatial covariates observed at each location  
 424 were used to model the coefficient of each basis function. This spatial dependence of the  
 425 coefficients enables the modelling of differences between the spatial locations. As our model  
 426 involves a multi-step fitting procedure, we developed a bootstrap approach for estimating  
 427 the standard errors of the model parameters. Our bootstrap approach involved resampling  
 428 blocks of data in both the temporal and spatial dimensions and was designed specifically  
 429 to account for the temporal and spatial relationships present in the data.

430 Recently, hierarchical Bayesian models have become a popular approach for analysing  
 431 spatio-temporal data. However, many such approaches involve deriving complex posterior  
 432 distributions and can be computationally intensive. Further, they require the model and all

parameters to be specified at the beginning. A key advantage of our procedure is that the model is fitted in a step-by-step fashion, which enables appropriate empirical choices to be made at each step. Our model also provides a very natural way of producing predictions, both for future time points and also at new locations. Once the model has been fitted, producing predictions at future time points only requires the extrapolation of the basis functions, and producing predictions at new spatial locations simply requires the spatial covariates at the new location to be known. Setting aside some of our data as test data, our model was able to predict monthly rainfall at future time points and new spatial locations relatively well.

An advantage of our approach to model building, compared to simply fitting a Bayesian model, is that the estimation of the covariance structure is separated out from the estimation of the regression structure. That is, the covariances can be investigated separately from the regression modelling. In addition, we note that the particular covariance structure will not affect the predictions produced by the model, as these are determined by the regression component of the model.

Ultimately, the goal of modelling and predicting rainfall in the MDB is to enhance understanding of rainfall patterns with the hope that this will ultimately aid in better water management. An important part of water management is also understanding how rainfall fluctuations affect river flow and surface runoff. Many of the currently used rainfall-runoff models are deterministic in nature. A potential future extension of this work is to incorporate our proposed model for rainfall in the MDB to develop stochastic methods for modelling the relationship between rainfall, river flow and runoff.

## Acknowledgements

This research was supported under Australian Research Council's Discovery Projects funding scheme (project number DP1092801).

## References

- Allcroft, D. J., Glasbey, C. A., 2003. A latent Gaussian Markov random-field model for spatiotemporal rainfall disaggregation. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 52, 487–498.
- Banerjee, S., Carlin, B. P., Gelfand, A. E., 2015. Hierarchical modeling and analysis for spatial data, 2nd Edition. CRC Press.
- Bogaert, P., Christakos, G., Jerrett, M., Yu, H.-L., 2009. Spatiotemporal modelling of ozone distribution in the State of California. *Atmospheric Environment* 43, 2471–2480.
- Carrera-Hernández, J. J., Gaskin, S. J., 2007. Spatio temporal analysis of daily precipitation and temperature in the Basin of Mexico. *Journal of Hydrology* 336, 231–249.
- Comas, C., Rodriguez-Cortes, F. J., Mateu, J., 2015. Second-order analysis of anisotropic spatiotemporal point process data. *Statistica Neerlandica* 69 (1), 49–66.
- Connell, D., Grafton, R. Q. (Eds.), 2011. Basin futures: water reform in the Murray-Darling basin. ANU E Press.
- Cressie, N., Wikle, C. K., 2011. Statistics for spatio-temporal data, 1st Edition. Wiley.
- Duan, N., 1983. Smearing estimate: A nonparametric retransformation method. *Journal of the American Statistical Association* 78 (383), 605–610.
- Eckert, N., Parent, E., Kies, R., Baya, H., 2010. A spatio-temporal modelling framework for assessing the fluctuations of avalanche occurrence resulting from climate change: application to 60 years of data in the northern French Alps. *Climatic Change* 101, 515–553.
- Feng, L., Nowak, G., O'Neill, T. J., Welsh, A. H., 2014. CUTOFF: A spatio-temporal imputation method. *Journal of Hydrology* 519, 3591–3605.

- 481 Fonseca, T. C. O., Steel, M. F. J., 2011. Non-Gaussian spatiotemporal modelling through  
482 scale mixing. *Biometrika* 98 (4), 761–774.
- 483 Fuentes, M., Guttorp, P., Sampson, P. D., 2006. Using transforms to analyze space-time  
484 processes. In: Finkenstadt, B., Held, L., Isham, V. (Eds.), *Statistical methods for spatio-*  
485 *temporal systems*. Chapman and Hall/CRC, Ch. 3, pp. 77–149.
- 486 Ghosh, S., Mallick, B. K., 2011. A hierarchical Bayesian spatio-temporal model for extreme  
487 precipitation events. *Environmetrics* 22 (2), 192–204.
- 488 Gryparis, A., Coull, B. A., Schwartz, J., Suh, H. H., 2007. Semiparametric latent variable  
489 regression models for spatiotemporal modelling of mobile source particles in the greater  
490 Boston area. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 56,  
491 183–209.
- 492 Holly, S., Pesaran, M. H., Yamagata, T., 2010. A spatio-temporal model of house prices in  
493 the USA. *Journal of Econometrics* 158, 160–173.
- 494 Lindstrom, J., Szpiro, A. A., Sampson, P. D., Sheppard, L., Oron, A., Richards, M., Larson,  
495 T., 2011. A flexible spatio-temporal model for air pollution: Allowing for spatio-temporal  
496 covariates. *UW Biostatistics Working Paper Series 370*, Department of Biostatistics,  
497 University of Washington, Seattle, WA.
- 498 Lovino, M., García, N. O., Baethgen, W., 2014. Spatiotemporal analysis of extreme pre-  
499 cipitation events in the Northeast region of Argentina (NEA). *Journal of Hydrology:*  
500 *Regional Studies* 2, 140–158.
- 501 Lowe, R., Bailey, T. C., Stephenson, D. B., Graham, R. J., Coelho, C. A. S., Carvalho,  
502 M. S., Barcellos, C., 2011. Spatio-temporal modelling of climate-sensitive disease risk:  
503 Towards an early warning system for dengue in Brazil. *Computers and Geosciences* 37,  
504 371–381.

- 505 Potter, N. J., Chiew, F. H. S., Frost, A. J., 2010. An assessment of the severity of recent  
506 reductions in rainfall and runoff in the Murray-Darling Basin. *Journal of Hydrology* 381,  
507 52–64.
- 508 Rodrigues, E. R., Gamerman, D., Tarumoto, M. H., Tzintzun, G., 2015. A non-  
509 homogeneous poisson model with spatial anisotropy applied to ozone data from Mexico  
510 City. *Environmental and Ecological Statistics* 22 (2), 393–422.
- 511 Sigrist, F., Künsch, H. R., Stahel, W. A., 2012. A dynamic nonstationary spatio-temporal  
512 model for short term prediction of precipitation. *The Annals of Applied Statistics* 6,  
513 1452–1477.
- 514 Smith, I., Chandler, E., 2010. Refining rainfall projections for the Murray Darling Basin  
515 of south-east Australia—the effect of sampling model results based on performance.  
516 *Climatic Change* 102, 377–393.
- 517 Szpiro, A. A., Sampson, P. D., Sheppard, L., Lumley, T., Adar, S. D., Kaufman, J. D., 2010.  
518 Predicting intra-urban variation in air pollution concentrations with complex spatio-  
519 temporal dependencies. *Environmetrics* 21, 606–631.
- 520 Zhao, J., 2015. The Anisotropic Spatiotemporal Estimation of Housing Prices. *The Journal*  
521 *of Real Estate Finance and Economics* 50 (4), 484–516.